Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Social Good

Coarse-to-Fine Domain Incremental Learning with Attentive Distillation for Mining Footprint Segmentation in Multispectral Imagery

Alif Tri Handoyo[1*], **Vincent C.S. Lee**[2], **Rizka Widyarini Purwanto**[1], **Alex M. Lechner**[1,3], **Deanna Kemp**[4] and **Muhamad Risqi U. Saputra**[1*]
[1]Monash University, Indonesia
[2]Monash University, Australia
[3]Northeastern University, China
[4]The University of Queensland, Australia
{alif.handoyo, vincent.cs.lee, rizka.purwanto, alex.lechner}@monash.edu, d.kemp@uq.edu.au, risqi.saputra@monash.edu

Abstract

Automatically mapping and segmenting global mining footprints using remote sensing and deep learning is critical for monitoring the socio-environmental risks and impacts of mining, yet its progress is hindered by the scarcity of fine-grained annotated data. Although large-scale datasets with coarse boundaries are widely available, leveraging them to improve fine-grained segmentation is challenging due to significant domain shift. To address this, we propose MineC2FNet, a coarse-to-fine domain incremental learning framework that exploits abundant coarse data to enhance fine-grained mining footprint segmentation. MineC2FNet adopts a teacher–student architecture with attentive distillation at both the feature and prediction levels, selectively transferring generalized knowledge from the coarse domain while enabling boundary refinement using limited fine-grained data (fine domain). We further introduce an expertly validated dataset of 219 images with precise boundary annotations across diverse geographies and commodities. Extensive experiments against state-of-the-art approaches, including domain adaptation and domain incremental learning methods, demonstrate that MineC2FNet achieves superior performance while effectively handling domain shift. The dataset and code are publicly available at https://github.com/risqiutama/MineC2FNet.

1 Introduction

Mapping Global Mining Footprints: A Socio-Environmental Monitoring Challenge. Mining is fundamental to the global economy, yet mineral resource extraction can cause profound, often irreversible damage to people and the environment [Owen *et al.*, 2023]. [1]

Extractive activities can lead to large-scale landscape alteration, ecosystem loss, biodiversity decline, and air and water pollution [Sonter *et al.*, 2018; Lechner *et al.*, 2017]. Many of these impacts can displace local and downstream populations and adversely affect the health of local communities and indigenous peoples [Ang *et al.*, 2023; Owen *et al.*, 2023]. These pressures and impacts can also drive resource-related conflicts and grievances.

Despite the promise of local economic development, host communities often live with long-term environmental legacies, long after mines cease operations [Kemp and Owen, 2025]. Multispectral remote sensing offers a method for continuous and repeatable tracking of mining footprints, in particularly in remote areas [Lechner *et al.*, 2019; Maus *et al.*, 2020; Saputra *et al.*, 2025]. This would enable scalable, remote monitoring of mining activity and its surrounding context, including how mining footprints change over time, and where change coincides with complex social and environmental risk factors.

However, a critical bottleneck remains: while Deep Learning (DL)-based semantic segmentation offers a transformative alternative to labor-intensive manual mapping [Wieland *et al.*, 2023; Cai *et al.*, 2023; Zhu *et al.*, 2025], these models depend on vast amounts of meticulously annotated data. Generating such data is exceptionally challenging, as distinguishing fine-grained features (e.g., pits, waste rock, tailings dams) requires substantial expert knowledge and field surveys [Maus *et al.*, 2020]. Although coarsely annotated data [Maus *et al.*, 2022] offers a cost-effective alternative with extensive coverage, it merges distinct land cover features into imprecise footprints (Figure 1 (a) - Task 1). Models trained on such labels inherit these imprecisions, leading to inaccurate boundary delineation and hindering effective environmental monitoring.

Our Contribution. To leverage both abundant coarse dataset and scarce fine dataset for mining footprint segmentation, and to bridge their domain gap in incremental learning settings, we propose Mine Coarse-to-Fine Net (MineC2FNet). MineC2FNet uses *attentive distillation* in a teacher–student training settings to adapt the model to new domains while selectively retaining knowledge from previous

*Corresponding authors.
[1]Supplementary material: http://arxiv.org/abs/2605.24460

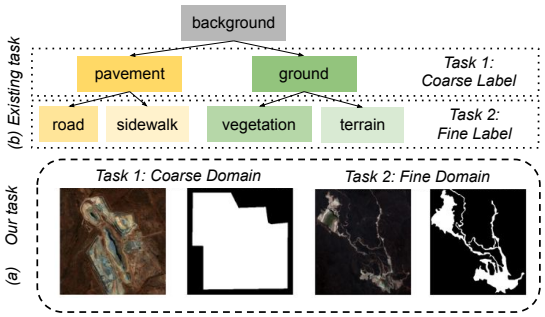

Figure 1: Comparison between (a) our coarse-to-fine problem with (b) existing coarse-to-fine continual learning task.

domains. The attentive distillation is applied at both feature (via *attentive feature injection*) and prediction levels (via *attentive knowledge transfer*) to transfer only high-quality delineation from the teacher (previous domain) to the student (new domain). Unlike prior coarse-to-fine continual learning methods defining "coarse" as broad label categories [Shenaj *et al.*, 2022; Alfarra *et al.*, 2024], we introduce a new task where "coarse" refers to imprecise boundaries and "fine" to accurate, expert-validated ones. The contributions of this paper are:

- We present MineC2FNet, the first coarse-to-fine domain incremental learning framework for mining footprint segmentation using multispectral imagery.
- We propose attentive distillation, a novel approach applying selective distillation to both feature and prediction levels.
- We introduce a new, expertly validated dataset of 219 globally distributed images for fine-grained mining footprint segmentation.
- We demonstrate the effectiveness of our framework and attentive distillation through comprehensive evaluation on this new dataset.

Team. Our project brought together a multidisciplinary team in remote sensing, deep learning, and mining domain experts, including professionals from the Faculty of IT at Monash University, the College of Resources and Civil Engineering at Northeastern University, China, and the Centre for Social Responsibility in Mining at the University of Queensland. These experts were key to guiding the project conceptualisation, identifying detailed mining features, and validating the quality of our new dataset. More broadly, funded by the Ford Foundation, our research contributes to the foundations mission of advancing social and environmental justice by improving transparency and accountability in mining.

2 Related Works

Semantic Segmentation of Mining Footprint. Applying deep learning-based semantic segmentation to multispectral satellite imagery has the potential to transform global mining footprint mapping and monitoring. Currently, global-scale studies, e.g., [Owen *et al.*, 2023], rely on a single geographic coordinate to represent mining operations, failing to capture their true spatial extent. On the other hand, studies that leverage supervised semantic segmentation are typically mine- or region-specific [Qiao *et al.*, 2024; Tong *et al.*, 2020], restricting their application to global datasets. Although [Saputra *et al.*, 2025] has demonstrated mining segmentation on a global dataset, their model was trained on only 37 locations, limiting its generalization capability. Finally, the success of these methods is fundamentally dependent on the availability of large-scale, pixel-level annotated datasets, which are unfortunately not yet available.

Continual Learning in Remote Sensing. Continual Learning (CL) addresses the challenge of training a model on a continuous stream of data or a sequence of tasks without experiencing *catastrophic forgetting*, a condition in which the model tends to lose previously learned knowledge when trained on a new dataset or tasks.

In remote sensing, the coarse-to-fine paradigm is applied to address CL problems in several distinct ways, ranging from data quality to architectural precision. To mitigate the problem of high annotation costs, some frameworks leverage coarse labels as a starting point to automatically generate more detailed, fine-grained pseudo-labels. These new pseudo labels then serve as the actual supervision to train the model. This strategy can be applied spatially, by turning imprecise polygons into sharp object boundaries [Das *et al.*, 2023], or semantically, by using coarse labels (broad classes, e.g., cropland) to identify and classify fine labels (specific sub-classes, e.g., paddy field) [Chen *et al.*, 2024]. Similarly, [Shenaj *et al.*, 2022] facilitates this coarse-to-fine CL by using knowledge distillation and a specialized weight initialization rule, allowing the model to learn new classes and adapt to new domains without being fully retrained.

Beyond class-incremental shifts, domain-incremental learning (DIL) tackles geographical variations by adapting the models to new spectral and spatial distributions without requiring access to original training data. The GSMF-RS-DIL [Huang *et al.*, 2024] framework utilizes graph space transformations and multifeature constraints to balance prior knowledge retention with the acquisition of new domain information. Similarly, multi-domain incremental architectures [Garg *et al.*, 2022] employ domain-specific parameters and differential learning rates within Domain-Aware Residual Units to optimize the stability-plasticity trade-off. These integrated strategies allow remote sensing models to remain effective across diverse environments while significantly mitigating catastrophic forgetting.

3 Data

Coarse Dataset. The coarse dataset was derived from the global-scale mining polygons (Version 2) data introduced by [Maus *et al.*, 2022], which originally contained over 44,929 mining polygons. This dataset provides a broader set of image features for mine footprint identification, characterized by rough, less precise edge delineations. These polygons are labeled as "Mining" and areas outside the polygons are considered "Non-Mining", resulting in a binary segmentation. To create a usable subset for our training, we filtered these polygons by establishing a minimum area threshold of ap-

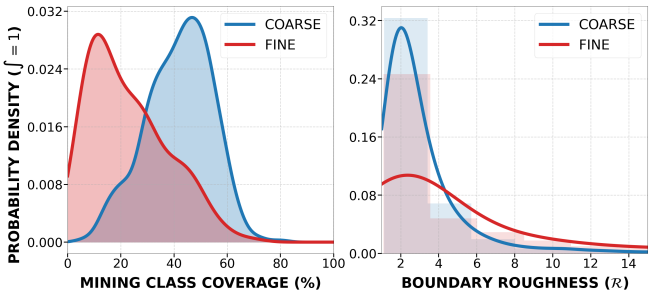

Figure 2: Domain shift analysis using Kernel Density Estimation (KDE). (Left) The coarse dataset exhibits a balanced class distribution, in contrast to the severe class imbalance of the fine dataset. (Right) Boundary roughness analysis demonstrating higher geometric complexity of fine masks compared to coarse masks.

proximately 9.8 km², corresponding to the smallest mine in our fine-grained dataset. Polygons with an area greater than this threshold were retained, resulting in a collection of 1,380 coarse-labeled images. Satellite imagery for these locations were sourced from Sentinel-2 (10m) and Landsat 8 (15/30m) satellites and included RGB, Near-Infrared (NIR), and Short-Wave Infrared (SWIR) bands.

Fine Dataset. The fine dataset is a high-quality, expert-validated collection of images featuring precise and detailed boundaries for mining segmentation. This dataset consists of 219 images, which include RGB and NIR bands, with a subset of 200 images also containing SWIR bands. These data were manually harmonised from the original datasets of [Werner *et al.*, 2020; Lechner *et al.*, 2016; Lechner *et al.*, 2019], following methods outlined in [Saputra *et al.*, 2025]. Similar to the coarse dataset, the resolution and extent of the multispectral data varied as we used Sentinel and Landsat 8. We split the dataset into training, validation, and test sets using an approximate 70:10:20 ratio. These locations was carefully balanced to account for a wide range of mine commodities, geographic regions, and climates. A key characteristic is its significant class imbalance, with non-mining pixels making up more than 80% of the labels across the dataset.

Domain Shift. Figure 2 demonstrates a clear domain gap between the coarse and fine datasets. The mining class coverage analysis (left) reveals a significant disparity in class labels pixel distributions: Coarse annotations exhibit a relatively balanced occupancy peaking at 45%, whereas the fine dataset reflects the extreme sparsity of precise footprints, peaking at only 12%. The Boundary Roughness (right) measured as $\mathcal{R} = \frac{P^2}{4\pi A}$ further differentiates the masks roughness. Coarse masks peak at $\mathcal{R} \approx 2$, identifying them as smooth, simplified blobs. In contrast, fine labels capture the irregular, high-frequency boundary details of real-world mines, with values reaching up to $\mathcal{R} \approx 12$. Collectively, these metrics confirm that while coarse labels provide a simplified approximation, fine labels capture the true geometric complexity and natural class imbalance inherent in the target domain.

4 Methodology

Our proposed method addresses the domain shift between coarse and fine datasets in training domain incremental learning model for mining footprint segmentation using multispectral satellite imagery. Figure 3 illustrates the overall workflow, highlighting our framework with attentive distillation applied at both the feature level (via Attentive Feature Injection) and the prediction level (via Attentive Knowledge Transfer). The core objective is to enable the model to learn from the new domain while selectively retaining knowledge from the previous domain.

4.1 Problem Formulation

We frame our coarse-to-fine segmentation problem as a domain incremental learning problem consisting of two sequential tasks. The goal is to train a model M, parameterized by θ, that leverages knowledge from an initial task trained on coarse domain D_c, to excel at a subsequent task with fine target domain D_f. This requires a domain incremental learning approach with selective attentive distillation, where the model is capable of transfering relevant general knowledge while discarding domain-specific noise (e.g., imprecise boundaries) from the coarse task.

Let $D_c = \{(X_i^c, Y_i^c)\}_{i=1}^{N_c}$ be the coarsely annotated dataset where X_i^c is an input image, Y_i^c is its corresponding coarse segmentation mask or label, and N_c is the number of coarse training images. Let $D_f = \{(X_i^f, Y_i^f)\}_{i=1}^{N_f}$ be the fine-grained, expertly annotated dataset, where Y_i^f is the precise, fine-grained mask, and N_f is the total images in the fine dataset. Our model, $M(\theta)$, which consists of a teacher component $M(\theta_T)$ and a student component $M(\theta_S)$, where $\theta = \{\theta_T, \theta_S\}$, will be trained using the following steps:

1. **Task 1 (Coarse Domain):** The teacher model $M(\theta_T)$ is trained on the coarse dataset D_c to learn a set of generalized parameters θ_T.
2. **Task 2 (Fine Domain):** With the teacher $M(\theta_T)$ frozen, the student model $M(\theta_S)$ is trained on the fine dataset D_f to learn its parameters θ_S. During this phase, the student must learn to predict the fine labels Y_f while selectively distilling knowledge from $M(\theta_T)$.

The final objective is to optimize the parameters θ_S as described in Equation 1, where L is the loss function.

$$M(\theta_S) = \arg\min_{\theta_S} L(\theta_S | D_f, M(\theta_T)) \quad (1)$$

4.2 Coarse-to-fine Domain Incremental Learning

In the context of our coarse-to-fine domain incremental learning, the training for Task 1 follows the standard training procedure for semantic segmentation, utilizing Binary Cross-Entropy (BCE) [Mao *et al.*, 2023] as the loss function. On the other hand, the training process for Task 2 is orchestrated by a composite loss function, L_{total}. This loss, shown in Eq. (2), is the core of our domain incremental learning approach. It combines L_{student}, the student's loss function on the fine data, with L_{distill}, the attentive distillation loss from the teacher.

$$L_{\text{total}}(D_f) = L_{\text{student}} + L_{\text{distill}} \quad (2)$$

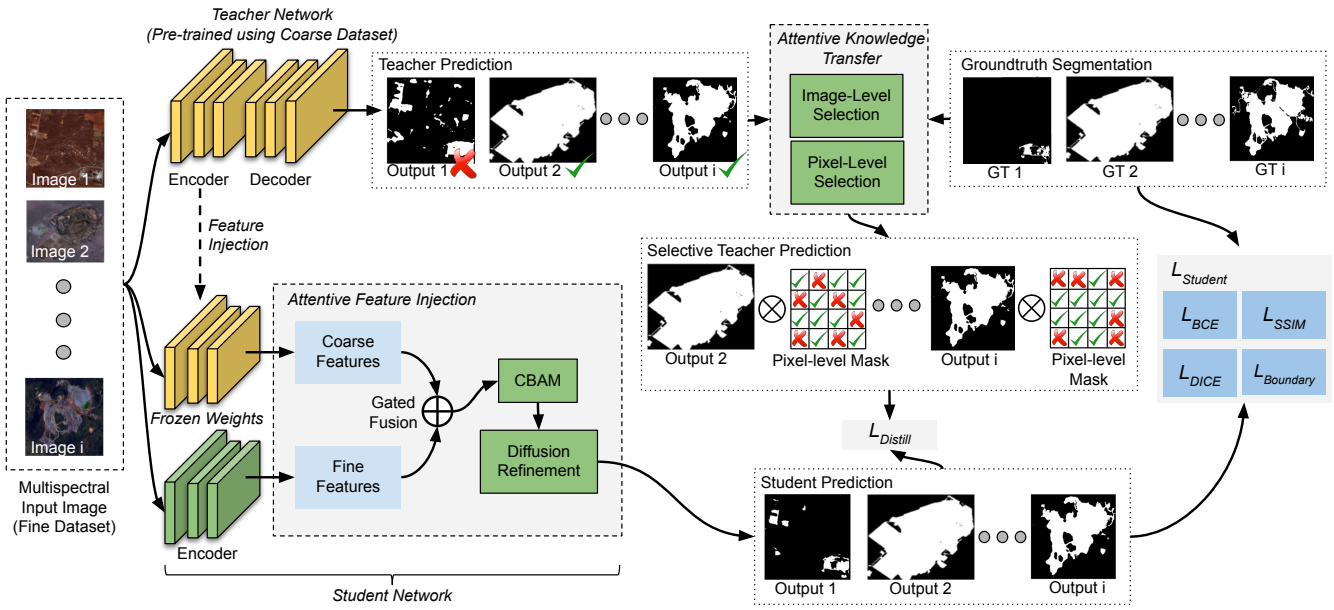

Figure 3: Overview of our coarse-to-fine domain incremental learning framework for mining footprint segmentation (MineC2FNet).

Student Loss. The student loss, L_{student}, is a comprehensive objective function designed to train the student model $M(\theta_S)$ on the fine-grained dataset D_f. As seen in Eq. (3), it consists of four components, including Binary Cross-Entropy (L_{BCE}), Dice Loss (L_{Dice}), Structural Similarity Loss (L_{SSIM}), and Boundary Loss (L_{Boundary}) functions.

$$L_{\text{student}}(D_f) = L_{\text{BCE}} + L_{\text{Dice}} + L_{\text{SSIM}} + L_{\text{Boundary}} \quad (3)$$

L_{BCE} is utilized as a widely used loss function for binary classification tasks, which we apply on a pixel-wise basis to measure the difference between the student's prediction $\hat{Y}_i^f$ and the ground truth fine mask Y_i^f. The Dice loss is added since it is particularly effective for segmentation tasks with imbalanced classes by maximizing the overlap between $\hat{Y}_i^f$ and Y_i^f [Sudre *et al.*, 2017]. Furthermore, we also incorporate L_{SSIM} [Zhao *et al.*, 2019] as this loss encourages the model to preserve the structural information of Y_i^f, which is crucial for maintaining realistic mining shapes.

Besides the loss functions above, we also incorporate L_{Boundary} to explicitly improve the model performance on edge delineation. This loss is designed by generating a boundary importance map from the ground truth mask and then using it to weight the L_{BCE}. In particular, we first apply Sobel filters to Y_i^f to compute the horizontal (G_x) and vertical (G_y) gradients. The boundary importance map $\hat{W} \in [0, 1]$ is then calculated as in Eq. (4).

$$\hat{W} = \frac{W - W_{\min}}{W_{\max} - W_{\min}}, \text{where } W = \sqrt{G_x^2 + G_y^2} \quad (4)$$

L_{Boundary} is then defined as a weighted binary-cross entropy loss as described in Eq. (5), where the weights are derived from image edges to emphasize optimizing the boundary regions of the segmented images.

$$L_{\text{Boundary}} = \frac{1}{K} \sum_{j=1}^{K} \hat{W}_j \cdot L_{\text{BCE}_j} \quad (5)$$

Note that K represents the number of pixel in the images. This formulation forces the model to pay more attention to correctly classifying pixels on the segmented boundaries.

Distillation Loss. We adopt a knowledge distillation where the loss (L_{Distill}) is formulated as the standard Kullback-Leibler (KL) divergence between the softened probability distributions of the teacher (P^T) and student (Q^T), scaled by a temperature factor T. As the teacher is trained in a domain with imperfect coarse labels, we introduce attentive distillation to ensure only reliable knowledge is transferred. This mechanism is applied at both the feature and prediction levels, as we will detail next.

4.3 Attentive Feature Injection

Our attentive feature injection is a multi-stage pipeline designed to intelligently transfer generalized feature extraction capabilities of the teacher model $M(\theta_T)$ to the student $M(\theta_S)$. This pipeline was designed to solve our domain incremental challenge with the understanding that $M(\theta_T)$ was trained on a larger set (source domain), capturing a more visually diverse range of global mining footprints. As a result, $M(\theta_T)$ possesses generalized knowledge of how mining areas appear in satellite imagery, even though its boundary delineation remains imprecise. To this end, rather than simply fine-tuning $M(\theta_T)$'s weight in the Task 2 or even concatenating $M(\theta_T)$'s and $M(\theta_S)$'s features, our attentive feature injection selectively fuses, refines, and enhances feature maps from both the teacher (generalized source domain features that uses coarse data) and student (specific target domain

features that uses fine data). As illustrated in Figure 3, the pipeline consists of four stages: Feature Injection, Gated Fusion (GF), Convolutional Block Attention Module (CBAM), and Diffusion Refinement (DR).

Feature Injection. A key distinction in our attentive distillation approach is that we inject a subset of $M(\theta_T)$, particularly the feature extractor, to $M(\theta_S)$, enabling $M(\theta_S)$ to incorporate dual, identical feature extractors: a "teacher feature extractor" (a frozen copy of the teacher's backbone, part of θ_T) and a "student feature extractor" (with trainable parameters, part of θ_S). During Task 2, for a given input image X_i^f, these two backbones operate in parallel. The student backbone extracts high-fidelity features directly from the fine-grained target domain input, while the frozen teacher backbone provides a parallel stream of generalized feature maps derived from the source domain. These two sets of features, one rich in precise target domain detail (F_{fine}) and the other in broad source domain context (F_{coarse}), are then passed to the subsequent attentive fusion and refinement stages.

Gated Fusion (GF). To dynamically merge contextual information from the coarse teacher backbone with high-frequency details from the fine student backbone, we employ Gated Fusion [Takikawa *et al.*, 2019]. This module's merit lies in its learnable, pixel-wise gating mechanism, which adaptively decides the contribution of each feature source. This allows the network to prioritize general context or precise detail as needed, creating a richer, fused representation without the need for fixed fusion rules.

CBAM. The fused features are then refined using the Convolutional Block Attention Module [Woo *et al.*, 2018]. We use CBAM for its efficiency and effectiveness in feature enhancement. Its primary benefit is that it sequentially infers attention maps along both channel and spatial dimensions. This process allows the model to learn what features are most salient (channel-wise) and where to focus its attention (spatial-wise), further boosting the discriminative power of the representations.

Diffusion Refinement (DR). As a final stage, we introduce a new Diffusion Refinement module to produce a highly polished representation for segmentation. Inspired by iterative refinement frameworks [Lin *et al.*, 2017a], our module progressively enhances features by propagating information through a sequence of custom refinement blocks. The key idea of our design is the structure of each block, which combines a convolution, a CBAM layer, and a residual connection to effectively sharpen noisy features. For each refinement step $t \in [1, T]$, the feature map F_{t-1} is updated using the operations detailed in Eq. (6)-(8).

$$F_{\text{conv}_t} = \text{Conv}_{3\times3}(F_{t-1}) \tag{6}$$

$$F_{\text{att}_t} = \text{CBAM}(F_{\text{conv}_t}) \tag{7}$$

$$F_t = F_{\text{att}_t} + F_{t-1} \tag{8}$$

This iterative process results in a final feature map, $F_{refined}$, with an enhanced focus on the most salient regions and channels, making it highly effective for achieving accurate semantic segmentation.

4.4 Attentive Knowledge Transfer

In addition to attentive feature injection, we introduce an attentive knowledge transfer mechanism at the prediction level. A key challenge is that a standard teacher trained on the coarse domain is not infallible; unconditional (non-selective) distillation transfers domain-specific noise like imprecise boundaries, degrading performance. To mitigate this, our framework adaptively selects *when* and *where* to apply the distillation loss, ensuring that the student only learns from the teacher's most confident and accurate predictions.

Image-level Selection. The first layer of our attentive mechanism operates at the image level, aiming to identify images within a training batch where the teacher model $M(\theta_T)$ produces more accurate predictions than the student model $M(\theta_S)$. This is based on the hypothesis that distillation is beneficial only when $M(\theta_T)$ is more accurate than $M(\theta_S)$ for a given input image. We formalize this by utilizing an indicator function $\mathbb{I}(\cdot)^{\text{img}}$ which returns 1 if $L_{\text{BCE}}(Y_i^f, P_i) < L_{\text{BCE}}(Y_i^f, Q_i)$ and 0 otherwise. Note that P_i and Q_i denote the output prediction (logits) of $M(\theta_T)$ and $M(\theta_S)$, respectively, for image i. The distillation loss with image-level selection is then formulated in Eq. (9).

$$L_{\text{distill}}^{\text{image}} = \frac{T^2}{N_f} \sum_{i=1}^{N_f} \mathbb{I}\begin{pmatrix} L_{\text{BCE}}(Y_i^f, P_i) \\ < L_{\text{BCE}}(Y_i^f, Q_i) \end{pmatrix}^{\text{img}} \cdot \text{KL}(P_i^T \parallel Q_i^T) \tag{9}$$

Pixel-level Selection. The second layer of attentive distillation applies at the pixel level. In the context of semantic segmentation, as L_{distill} will be computed for every single pixel, we realize that not all per-pixel teacher's prediction will be accurate, especially when the teacher was trained using imprecise ground truth labels. To this end, we selectively apply distillation only when the teacher's prediction at the pixel level is more accurate than the student's prediction. Similar to image-level selection, we formalize this pixel-level selection by utilizing an indicator function $\mathbb{I}(\cdot)^{\text{pxl}}$ which returns 1 if $L_{\text{BCE}}(Y_{ij}^f, P_{ij}) < L_{\text{BCE}}(Y_{ij}^f, Q_{ij})$ and 0 otherwise. Note that $j \in K$ represents the pixel index in image i. The distillation loss with pixel-level selection is formulated in Equation 10.

$$L_{\text{distill}}^{\text{pixel}} = \frac{T^2}{N_f} \sum_{i,j} \mathbb{I}\begin{pmatrix} L_{\text{BCE}}(Y_{ij}^f, P_{ij}) \\ < L_{\text{BCE}}(Y_{ij}^f, Q_{ij}) \end{pmatrix}^{\text{pxl}} \cdot \text{KL}(P_{ij}^T \parallel Q_{ij}^T) \tag{10}$$

Note that in practice, $\mathbb{I}^{\text{pxl}}$ will be in a matrix form rather than in a scalar form as in $\mathbb{I}^{\text{img}}$.

Hybrid Selection. Our final approach combines both image-level and pixel-level selection strategies into a hybrid selection mask $\mathbb{I}^{\text{hybrid}}$ by performing element-wise multiplication of $\mathbb{I}^{\text{pxl}}$ and $\mathbb{I}^{\text{img}}$ as shown in Eq. (11).

$$\mathbb{I}_{ij}^{\text{hybrid}} = \mathbb{I}_i^{\text{img}} \cdot \mathbb{I}_{ij}^{\text{pxl}} \tag{11}$$

This ensures that distillation is applied only to pixels where the teacher is more accurate than the student, and only within input images where the teacher demonstrates overall superior

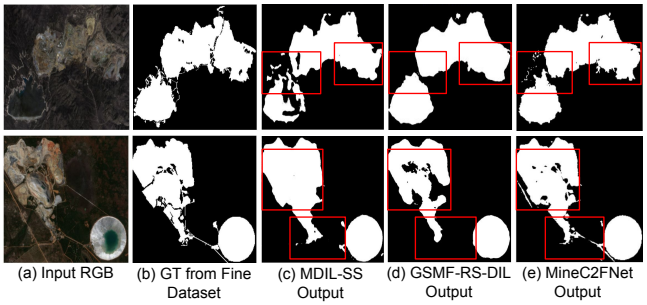

Figure 4: Segmentation results of MineC2FNet and several baseline models on the fine-grained test set.

performance. The final distillation loss is computed as defined in Equ. (12):

$$L_{\text{distill}} = \frac{\sum_{i,j} \left(\mathbb{I}_{ij}^{\text{hybrid}} \cdot T^2 \cdot \text{KL}(P_{ij}^T \parallel Q_{ij}^T) \right)}{\sum_{i,j} \mathbb{I}_{ij}^{\text{hybrid}} + \epsilon} \tag{12}$$

where the numerator accumulates the KL divergence values, and the denominator normalizes it via the total number of selected pixels (with ϵ for numerical stability). This hierarchical and attentive strategy ensures that the student selectively learns only from the teacher's most reliable predictions.

4.5 Training

Experimentation Setting. We implemented our framework using a Feature Pyramid Network (FPN) with a DenseNet-121 backbone, initialized with ImageNet pretrained weights and modified to process 6-channel inputs (RGB, NIR, SWIR) at a resolution of 256×256. The model was trained using the AdamW optimizer [Loshchilov and Hutter, 2019] with an initial learning rate of 1×10^{-3}, weight decay of 1×10^{-4}, and a batch size of 8 for a maximum of 100 epochs. Standard geometric augmentations, including random flips, rotations, and scaling, were applied to enhance generalization. All experiments were conducted on an NVIDIA RTX A4000 GPU.

Evaluation Metrics and Baselines. To assess performance, we benchmark MineC2FNet against five distinct learning paradigms: Transfer Learning (including the foundation model, e.g., Prithvi), Domain Adaptation, Baseline Continual Learning, Class Incremental Learning (CIL), and Domain Incremental Learning (DIL). Performance is evaluated using three standard metrics: *Pixel Accuracy*, *Mean F1 Score (mF1)*, and *Mean Intersection over Union (mIoU)*. More details can be found in the supplementary material.

5 Results and Discussion

5.1 Overall Performance

Quantitative Analysis. Table 1 summarizes the performance of MineC2FNet against state-of-the-art methods. Our model, termed MineC2FNet, demonstrates superior performance, achieving the highest Accuracy (92.33%) and mIoU (73.64%) compared to different methods across five distinct learning categories. It is worth noting that while leading DIL methods like MDIL-SS achieve a slightly higher mF1 score (85.36% vs. 84.10%), our framework secures a clear advantage in mIoU, surpassing them by over 1.6 percentage points. This numerical superiority translates directly to visual precision, as shown in Figure 4, in which MineC2FNet successfully captures more detailed boundaries compared to the existing DIL models. Conversely, specialized CIL and Domain Adaptation methods struggle with this specific coarse-to-fine domain shift. Notably, methods such as SPPA (31.85%) and BUS (42.35%) perform significantly worse than even standard baseline continual learning approaches like LwF or LwM. This confirms that standard class-focused or domain adaptation strategies are ill-suited for this problem, whereas our approach successfully produces fine details while preserving coarse general knowledge.

We further compare our method with the coarse baseline on the coarse-grained dataset, as reported in Table 2. Al-

Model	Acc.	mF1	mIoU
CNN and Transformer-based methods (Transfer Learning)			
U-Net [Ronneberger *et al.*, 2015]	89.79	79.52	67.26
DeepLabV3+ [Chen *et al.*, 2018]	90.01	79.57	67.47
FPN [Lin *et al.*, 2017b]	90.65	80.06	68.26
Prithvi [Jakubik *et al.*, 2023]	66.14	66.21	49.74
Domain Adaptation			
UDAforRS [Li *et al.*, 2022]	75.16	72.42	57.49
BUS [Choe *et al.*, 2024]	60.22	59.21	42.35
Baseline Continual Learning			
LwF [Li and Hoiem, 2018]	90.22	78.94	67.02
LwM [Dhar *et al.*, 2019]	90.60	79.36	67.43
Replay [Rolnick *et al.*, 2019]	88.81	76.50	63.61
Class Incremental Learning			
SPPA [Lin *et al.*, 2022]	54.68	48.32	31.85
CCDA [Shenaj *et al.*, 2022]	78.90	72.90	57.00
Domain Incremental Learning			
MDIL-SS [Garg *et al.*, 2022]	88.62	85.36	71.99
GSMF-RS-DIL [Huang *et al.*, 2024]	88.83	85.02	71.25
MineC2FNet (ours)	**92.33**	**84.10**	**73.64**

Table 1: Performance comparison of different models across various models and learning paradigms.

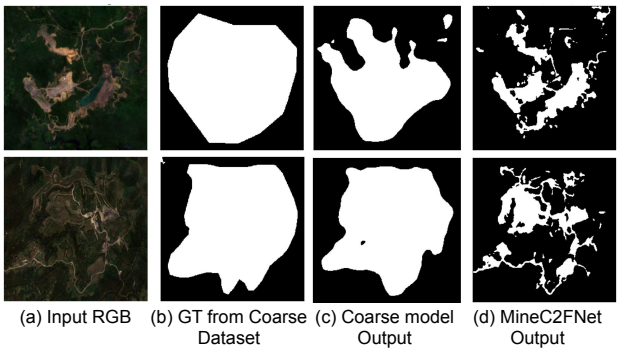

Figure 5: Through our framework, MineC2FNet produces more fine-grained output in the coarse dataset.

Metric	Coarse Baseline	MineC2FNet (Ours)
Accuracy	0.8733	0.7681
mF1	0.8331	0.5612
mIoU	0.7249	0.4429

Table 2: Performance comparison between the coarse baseline and the proposed MineC2FNet.

GF	CBAM	DR	Acc.	mF1	mIoU	Δ mIoU
✗	✗	✗	89.00	75.39	62.35	–
✓	✗	✗	90.31	76.72	64.35	+2.00
✓	✓	✗	**91.33**	79.15	67.79	+5.44
✓	✓	✓	91.08	**79.84**	**68.41**	**+6.06**

Table 3: Ablation study of Attentive Feature Injection.

though MineC2FNet achieves a lower mIoU (44.29%) than the coarse baseline (72.49%), this outcome is expected due to the imprecise annotations in the coarse dataset. The coarse baseline is trained and evaluated entirely on the source domain, and its predictions therefore closely match the coarse labels, which are known to contain inaccurate boundaries. In contrast, our model adapts to the target domain with finer and more precise annotations (see Figure 5), resulting in lower agreement when evaluated against coarse labels. This discrepancy does not indicate a failure of our approach; rather, it reflects superior boundary delineation while maintaining generalization by retaining useful knowledge from both coarse and fine domain. Evaluations across five Köppen–Geiger climate zones (Beck et al. 2018) confirmed the model's global robustness and generalizability (see supplementary material).

Qualitative Analysis. Figure 4 highlights the precise segmentation masks generated by MineC2FNet, demonstrating its ability to accurately delineate complex boundaries. As shown in the second row, our model reduces false positives caused by noisy features more effectively than the baselines. Furthermore, when tested on the coarse dataset, our domain-adapted MineC2FNet yields fine-grained segmentation that reflects true features on the ground (Figure 5(d)). In contrast, training the model only on the coarse dataset inherits imprecise identification and poor boundary delineation (Figure 5(c)), underscoring the importance of our coarse-to-fine domain incremental learning.

Model Complexity and Efficiency. Finally, we evaluate MineC2FNet efficiency against the FPN baseline. The model achieves a substantial 5.4 point mIoU improvement, with a negligible speed reduction (13.38 to 11.00 FPS) due to added attention mechanisms. We consider this trade-off justified by the performance gain (detailed results in supplementary material).

5.2 Ablation Study

Impact of Attentive Feature Injection. We conducted ablation studies to clearly isolate the effects of each module within our attentive feature injection mechanism, i.e., Gated Fusion (GF), CBAM, and Diffusion Refinement (DR). Note that only RGB and NIR bands were used for this experiment. As shown in Table 3, while adding GF alone can improves performance, incorporating CBAM and DR yields a significant performance gain. The highest mIoU (68.41%) is achieved when all three components are combined. This confirms that these modules contribute critically to the complete pipeline for optimal fusion of coarse and fine features.

Strategy	Acc.	mF1	mIoU	Δ mIoU
No Distill. (Base)	89.00	75.39	62.35	–
Standard Distill.	89.38	75.79	62.59	+0.24
Image-level Sel.	90.42	78.79	67.17	+4.82
Pixel-level Sel.	**91.14**	79.30	67.84	+5.49
Hybrid Sel.	91.10	**79.61**	**67.91**	**+5.56**

Table 4: Comparison of different Attentive Knowledge Transfer strategies (all configurations include feature injection).

Band Comb.	Acc.	mF1	mIoU	Δ mIoU
RGB (Baseline)	87.08	71.12	57.09	–
RGB+NIR	89.00	75.39	62.35	+5.26
RGB+NIR+SWIR	**89.68**	**79.29**	**67.29**	**+10.20**

Table 5: Impact of different band combinations.

Impact of Attentive Knowledge Transfer. Table 4 highlights the impact of different distillation methods. A model trained with standard distillation shows only marginal improvement over the baseline, confirming the risk of negative knowledge transfer. In contrast, our attentive distillation, at either image or pixel level, shows a large improvement (≥ 4.82 ΔmIoU). The hybrid method, distilling knowledge only from the teacher's most accurate regions, achieves the highest performance, validating that a selective, hierarchical mechanism is critical for successful coarse-to-fine domain knowledge transfer.

Bands Comparison. As shown in Table 5, incorporating non-visible spectral bands is critical. While adding NIR significantly increases performance, the optimal results are achieved using the full RGB-NIR-SWIR combination. This underscores the value of NIR and SWIR bands in identifying distinct mining footprint features [Saputra *et al.*, 2025].

6 Conclusion

In this paper, we addressed the critical challenge of domain shift between coarse (abundant) and fine (scarce) datasets in training domain incremental learning model for mining footprint segmentation. We introduce MineC2FNet, a novel coarse-to-fine domain incremental framework. Our core contribution, an attentive distillation mechanism, proved highly effective, achieving a state-of-the-art mIoU of 73.64% by successfully selectively transferring knowledge from coarse to fine-grained data and outperforming other continual learning models. This work also introduced a new, expertly validated global dataset to facilitate further research in this domain. However, given our current limitation to binary segmentation, future work will extend this framework to complex multi-class segmentation of diverse mining and non-mining categories.

Acknowledgments

This study was supported in part by the Google Research Scholar Program, Australian Research Council LP200301160, and the Ford Foundation. The work of A. T. H. was partially supported by the Monash University Indonesia PhD Tuition Scholarship. We also sincerely thank John R. Owen for his contribution to the development of the training data.

References

[Alfarra *et al.*, 2024] Motasem Alfarra, Zhipeng Cai, Adel Bibi, Bernard Ghanem, and Matthias Müller. Simcs: Simulation for domain incremental online continual segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):10795–10803, Mar. 2024.

[Ang *et al.*, 2023] Michelle Li Ern Ang, John R. Owen, Christopher N. Gibbins, Éléonore Lèbre, Deanna Kemp, Muhamad Risqi U. Saputra, Jo-Anne Everingham, and Alex M. Lechner. Systematic review of gis and remote sensing applications for assessing the socioeconomic impacts of mining. *The Journal of Environment & Development*, 32(3):243–273, 2023.

[Cai *et al.*, 2023] Zhiwen Cai, Haodong Wei, Qiong Hu, Wei Zhou, Xinyu Zhang, Wenjie Jin, Ling Wang, Shuxia Yu, Zhen Wang, Baodong Xu, et al. Learning spectral-spatial representations from vhr images for fine-scale crop type mapping: A case study of rice-crayfish field extraction in south china. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199:28–39, 2023.

[Chen *et al.*, 2018] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.

[Chen *et al.*, 2024] Hao Chen, Wen Yang, Li Liu, and Gui-Song Xia. Coarse-to-fine semantic segmentation of satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 217:1–17, 2024.

[Choe *et al.*, 2024] Seun-An Choe, Ah-Hyung Shin, Keon-Hee Park, Jinwoo Choi, and Gyeong-Moon Park. Open-set domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23943–23953, 2024.

[Das *et al.*, 2023] Anurag Das, Yongqin Xian, Yang He, Zeynep Akata, and Bernt Schiele. Urban scene semantic segmentation with low-cost coarse annotation. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5967–5976, 2023.

[Dhar *et al.*, 2019] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5138–5146, 2019.

[Garg *et al.*, 2022] Prachi Garg, Rohit Saluja, Vineeth N Balasubramanian, Chetan Arora, Anbumani Subramanian, and CV Jawahar. Multi-domain incremental learning for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 761–771, 2022.

[Huang *et al.*, 2024] Wubiao Huang, Mingtao Ding, and Fei Deng. Domain-incremental learning for remote sensing semantic segmentation with multifeature constraints in graph space. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

[Jakubik *et al.*, 2023] Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi, Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation models for generalist geospatial artificial intelligence, 2023.

[Kemp and Owen, 2025] Deanna Kemp and John R. Owen. Global mining and the production of inequality: a case for continued inquiry. *World Development Perspectives*, 39:100708, 2025.

[Lechner *et al.*, 2016] Alex Mark Lechner, Owen Kassulke, and Corinne Unger. Spatial assessment of open cut coal mining progressive rehabilitation to support the monitoring of rehabilitation liabilities. *Resources Policy*, 50:234–243, December 2016. Publisher Copyright: © 2016 Elsevier Ltd Copyright: Copyright 2017 Elsevier B.V., All rights reserved.

[Lechner *et al.*, 2017] Alex M. Lechner, Neil McIntyre, Katherine Witt, Christopher M. Raymond, Sven Arnold, Margaretha Scott, and Will Rifkin. Challenges of integrated modelling in mining regions to address social, environmental and economic impacts. *Environmental Modelling & Software*, 93:268–281, 2017.

[Lechner *et al.*, 2019] Alex M Lechner, John Owen, Michelle Li Ern Ang, Mansour Edraki, Nor Aklima Che Awang, and Deanna Kemp. Historical socio-environmental assessment of resource development footprints using remote sensing. *Remote Sensing Applications: Society and Environment*, 15:100236, 2019.

[Li and Hoiem, 2018] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

[Li *et al.*, 2022] Weitao Li, Hui Gao, Yi Su, and Biffon Manyura Momanyi. Unsupervised domain adaptation for remote sensing semantic segmentation with transformer. *Remote Sensing*, 14(19), 2022.

[Lin *et al.*, 2017a] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *2017*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5168–5177, 2017.

[Lin *et al.*, 2017b] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[Lin *et al.*, 2022] Zihan Lin, Zilei Wang, and Yixin Zhang. Continual semantic segmentation via structure preserving and projected feature alignment. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 345–361, Cham, 2022. Springer Nature Switzerland.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[Mao *et al.*, 2023] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications, 2023.

[Maus *et al.*, 2020] Victor Maus, Stefan Giljum, Jakob Gutschlhofer, Dieison M. da Silva, Michael Probst, Sidnei L. B. Gass, Sebastian Luckeneder, Mirko Lieber, and Ian McCallum. A global-scale data set of mining areas. *Scientific Data*, 7(1):289, Sep 2020.

[Maus *et al.*, 2022] Victor Maus, Dieison M da Silva, Jakob Gutschlhofer, Robson da Rosa, Stefan Giljum, Sidnei L B Gass, Sebastian Luckeneder, Mirko Lieber, and Ian McCallum. Global-scale mining polygons (Version 2), 2022.

[Owen *et al.*, 2023] John R. Owen, Deanna Kemp, Alex M. Lechner, Jill Harris, Ruilian Zhang, and Éléonore Lèbre. Energy transition minerals and their intersection with land-connected peoples. *Nature Sustainability*, 6(2):203–211, Feb 2023.

[Qiao *et al.*, 2024] Qinghua Qiao, Yanyue Li, and Huaquan Lv. Open-pit mining area extraction using multispectral remote sensing images: A deep learning extraction method based on transformer. *Applied Sciences*, 14(14), 2024.

[Rolnick *et al.*, 2019] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. Experience replay for continual learning, 2019.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015.

[Saputra *et al.*, 2025] Muhamad Risqi U Saputra, Irfan Dwiki Bhaswara, Bahrul Ilmi Nasution, Michelle Ang Li Ern, Nur Laily Romadhotul Husna, Tahjudil Witra, Vicky Feliren, John R Owen, Deanna Kemp, and Alex M Lechner. Multi-modal deep learning approaches to semantic segmentation of mining footprints with multispectral satellite imagery. *Remote Sensing of Environment*, 318:114584, 2025.

[Shenaj *et al.*, 2022] Donald Shenaj, Francesco Barbato, Umberto Michieli, and Pietro Zanuttigh. Continual coarse-to-fine domain adaptation in semantic segmentation. *Image and Vision Computing*, 121:104426, 2022.

[Sonter *et al.*, 2018] Laura J Sonter, Saleem H Ali, and James E M Watson. Mining and biodiversity: key issues and research needs in conservation science. *Proc Biol Sci*, 285(1892), December 2018.

[Sudre *et al.*, 2017] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*, page 240–248. Springer International Publishing, 2017.

[Takikawa *et al.*, 2019] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation, 2019.

[Tong *et al.*, 2020] Xin-Yi Tong, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237:111322, 2020.

[Werner *et al.*, 2020] Tim T Werner, Gavin M Mudd, Aafke M Schipper, Mark AJ Huijbregts, Lakshay Taneja, and Stephen A Northey. Global-scale remote sensing of mine areas and analysis of factors explaining their extent. *Global Environmental Change*, 60:102007, 2020.

[Wieland *et al.*, 2023] Marc Wieland, Sandro Martinis, Ralph Kiefl, and Veronika Gstaiger. Semantic segmentation of water bodies in very high-resolution satellite and aerial images. *Remote Sensing of Environment*, 287:113452, 2023.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, page 3–19, Berlin, Heidelberg, 2018. Springer-Verlag.

[Zhao *et al.*, 2019] Shuai Zhao, Boxi Wu, Wenqing Chu, Yao Hu, and Deng Cai. Correlation maximized structural similarity loss for semantic segmentation, 2019.

[Zhu *et al.*, 2025] Xiaoqian Zhu, Xiangrong Zhang, Tianyang Zhang, Chaowei Fang, Xu Tang, and Licheng Jiao. Regionmatch: Pixel-region collaboration for semi-supervised semantic segmentation in remote sensing images. In James Kwok, editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 2530–2538. International Joint Conferences on Artificial Intelligence Organization, 8 2025. Main Track.